

DZero Data Processing and Monte Carlo Simulation

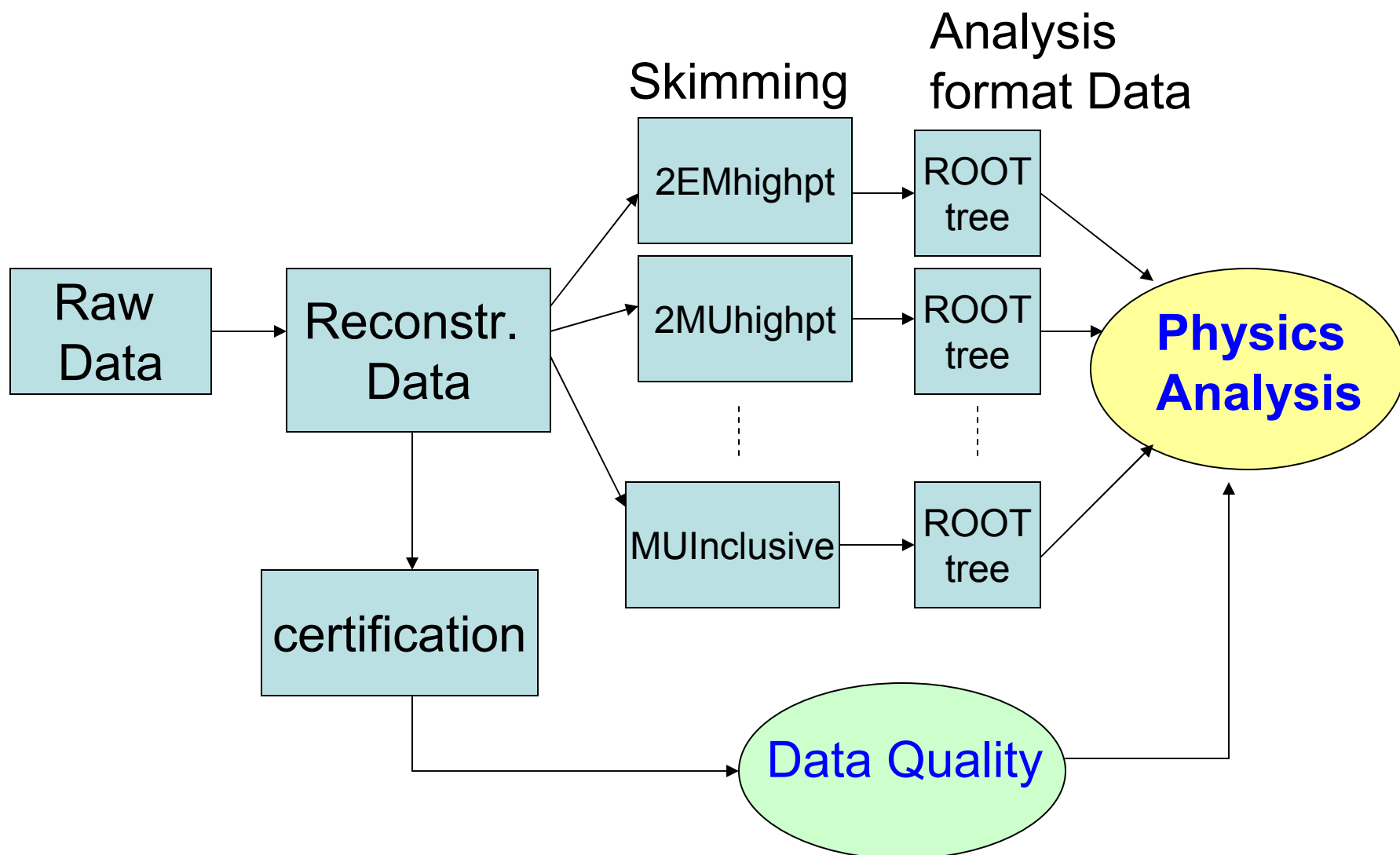
Qizhong Li

Fermilab

Oct. 6, 2008

All Experimenters' Meeting

Data Processing



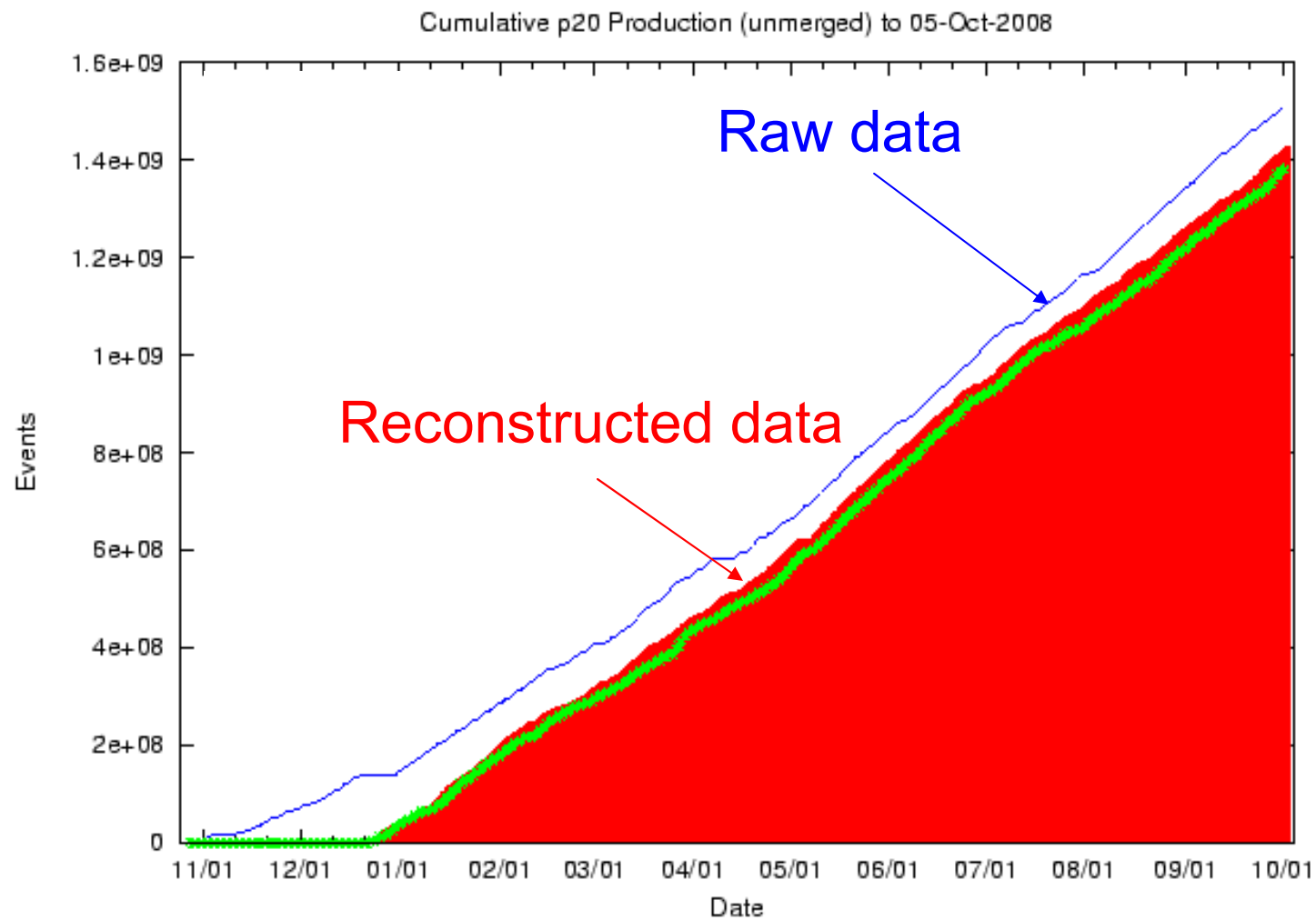
Data size

- Total storage (as today): 3808 TB.
- Raw data: Total 4.65 Billion events in Run II;
- File size:
 - Raw data: 200KB/event;
 - Reconstructed data: 120KB/event;
 - Analysis format: 75KB/event;

Raw Data Processing Strategy

- Process raw data as soon as we can.
- Wait 2 days after data taking to allow for calibration constants to be determined and propagated to offline database.
- Try to keep as small a backlog as possible.
- Be responsive to special needs:
 - Detector configuration changes
 - Special runs
 - Whenever offline feedback is needed quickly

Raw Data Processed

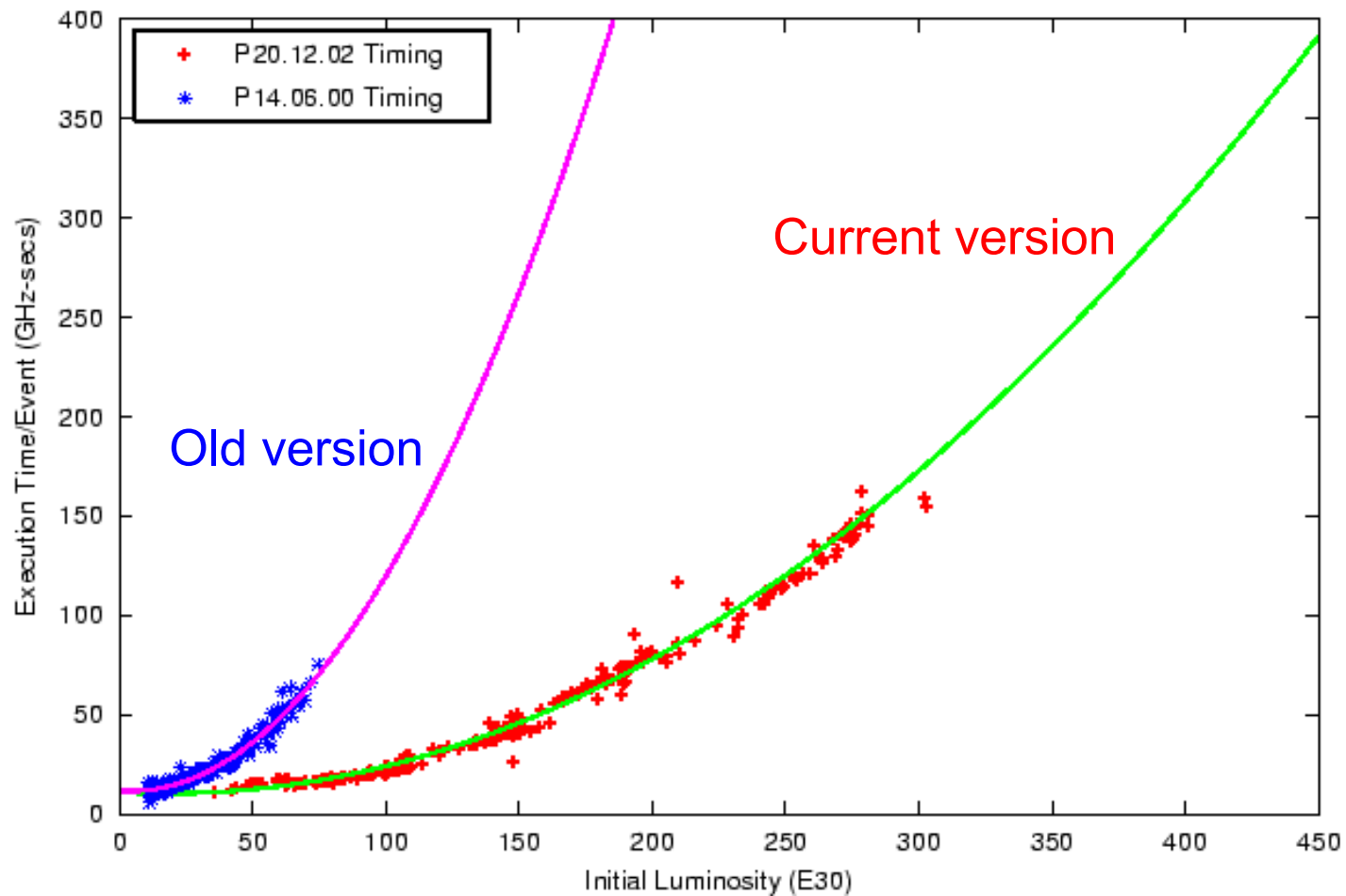


Processing of Raw Data

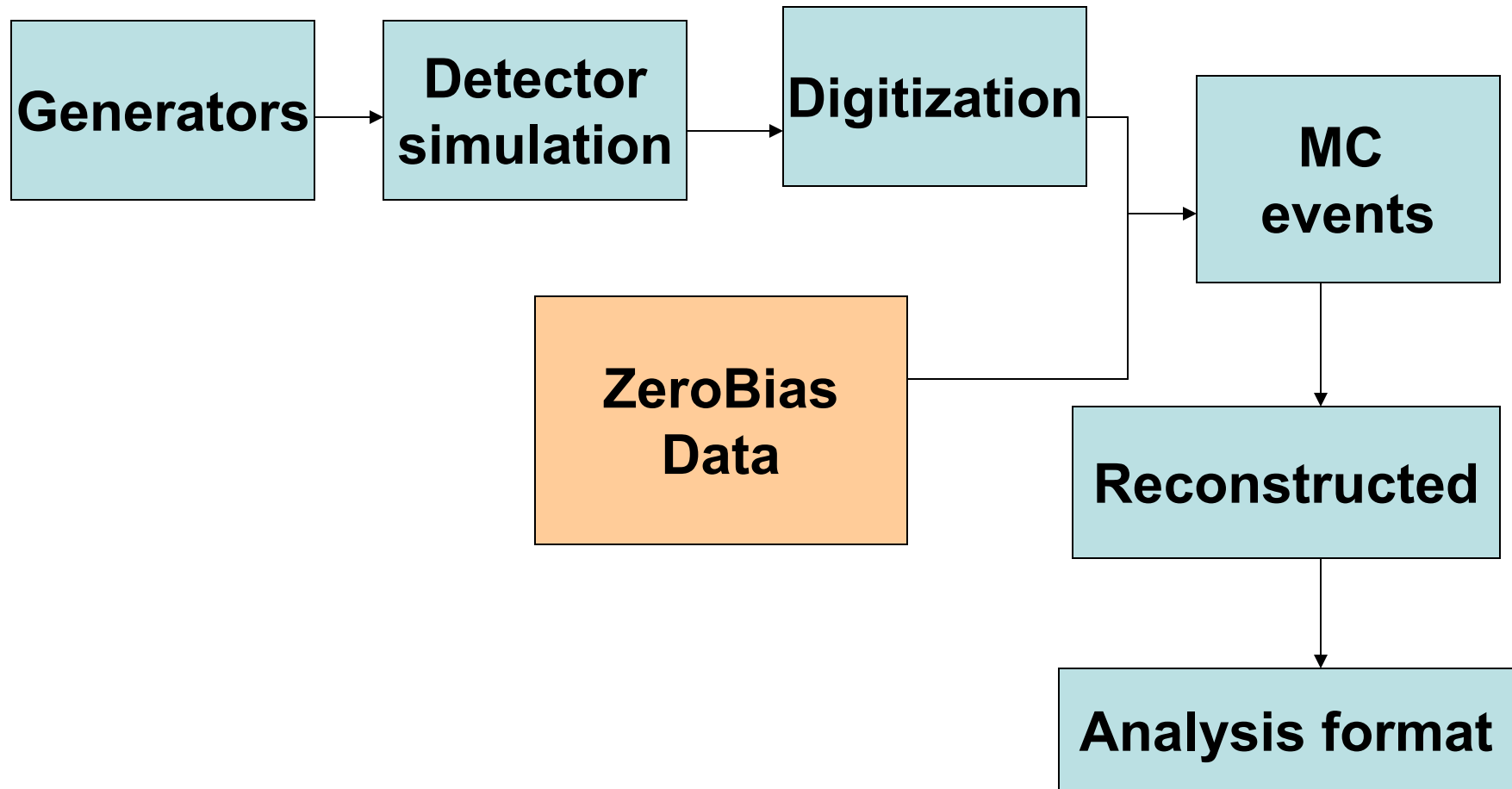
- Data reconstruction is done on D0 computer cluster (part of FermiGrid), using SamGrid system.
- Average production rate (12/2007-08/2008 average):
 - Raw data: 4.6 M events/day
 - Data processed: 5.1 M events/day
- Last week production:
 - Raw data collected: 30.1 M events;
 - Data processed: 36.1 M events.
- Current backlog: ~3 weeks.
 - CD has formed a task force to improve the efficiency lost in SamGrid.
 - Before SamGrid problem is solved, temporarily moved more CPUs from analysis to data processing.

Reco Timing

D0 Reconstruction CPU Timing vs Initial Luminosity

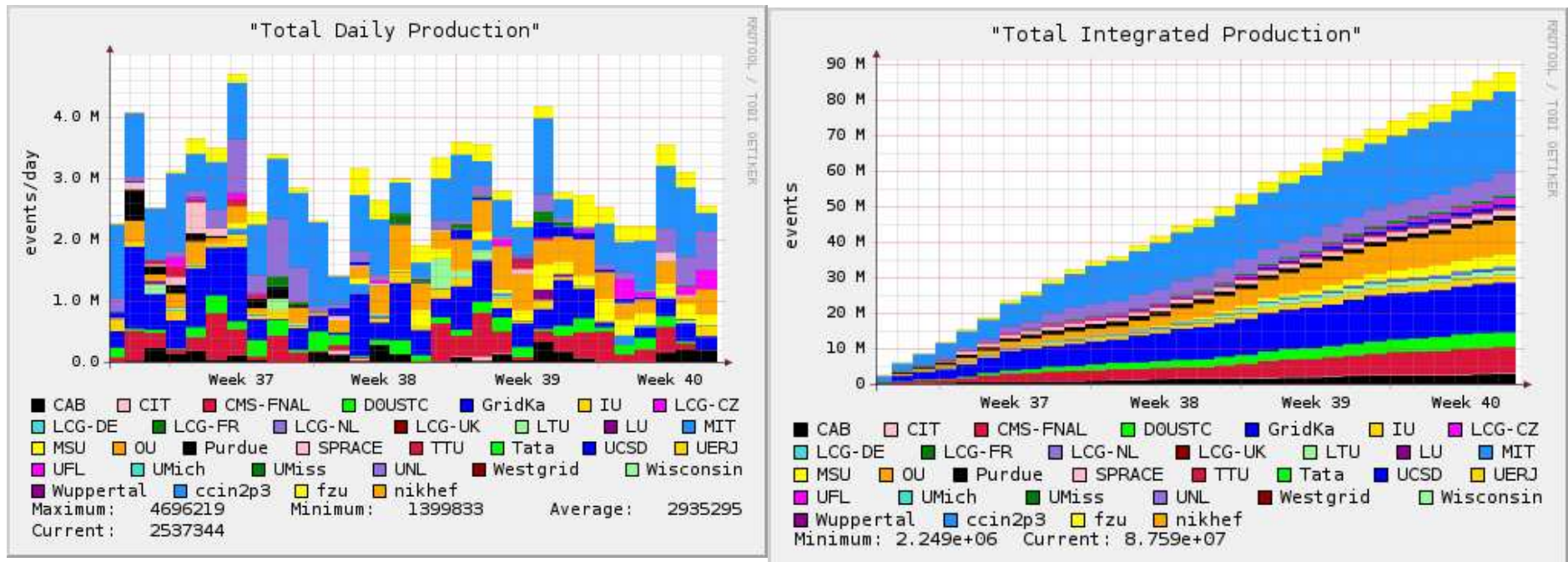


Monte Carlo Generation



Monte Carlo Generation

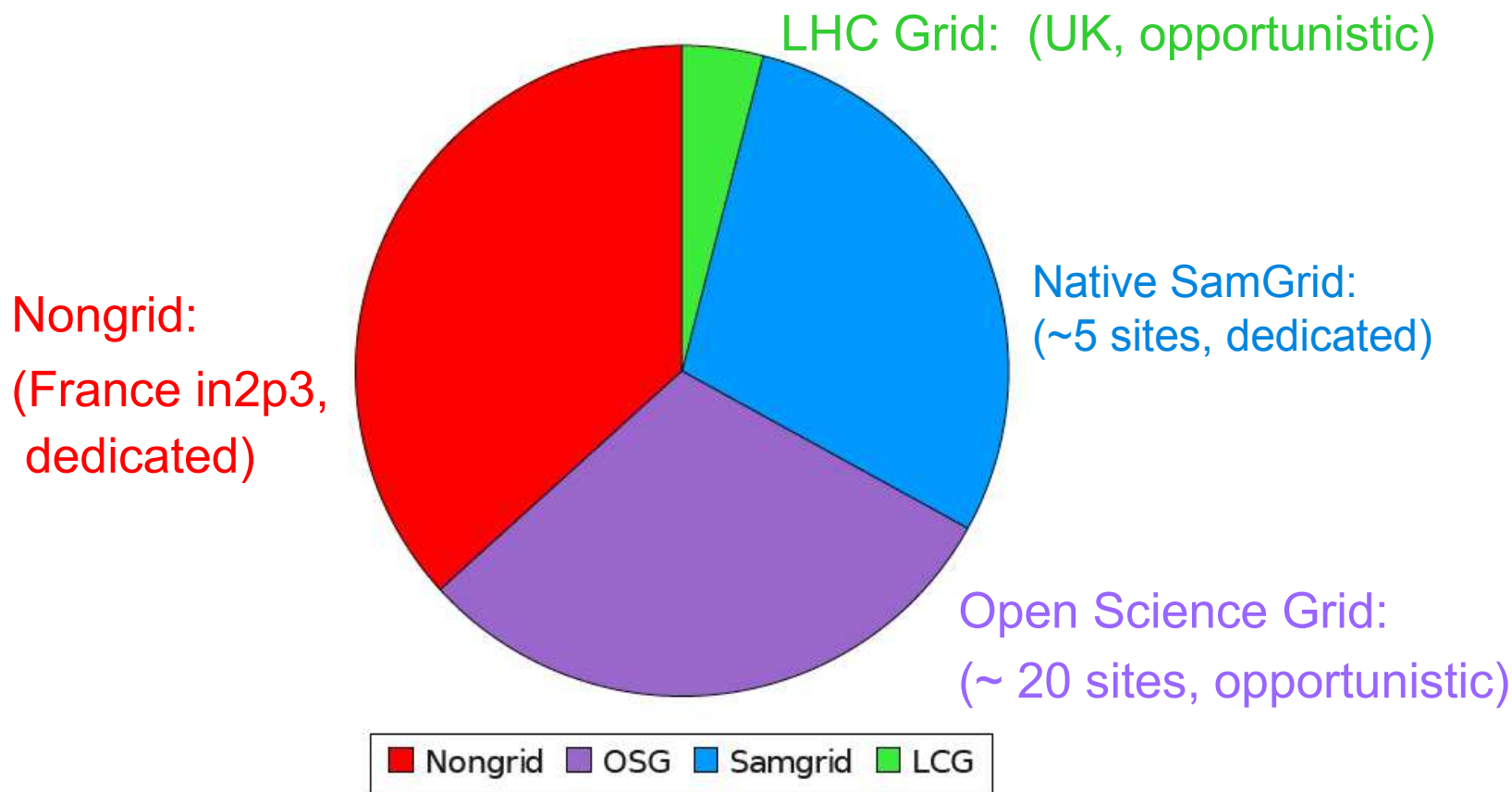
- Monte Carlo events are mainly generated at remote sites, most of them are on grid.



Averaging 2.94M per day and in total 87.6M in last 30 days

MC Generation for Past Year

D0 runs MC generation on a variety of dedicated and opportunistic grid resources (~ 700 M MC events last year)



MC Generation Last week

- Last week MC generation:
 - Total: 18.2 M events;
 - NonGrid (CCin2p3): 4.4 M events;
 - Open Science Grid sites: 9.9 M events;
 - Native SamGrid: 3.9 M events.
- OSG efficiency has big improvement in the recent weeks.
 - From average (9/2007-8/2008): 4.2 M events / week
 - To 11.0 M events (a record high) the week before last.
(Thanks to OSG tech team!)

Analysis Computing

- Two analysis computing clusters:
 - CAB (Central Analysis Backend):
 - Total ~5000 CPUs
(3400 CPUs for analysis; 1640 CPUs for data processing).
 - CAB is managed by Computing Division.
 - Clued0:
 - Total: ~540 CPUs
 - Managed by the D0 collaborating institutions.
 - Both CAB and Clued0 provide reliable and efficient performance.
- CPU intensive jobs (like Matrix Element Analysis) are using grid at remote sites.

Summary

- D0 raw data processing following data taking as soon as possible.
 - Provide reconstructed data for analysis in a timely manner
 - Provide fast turnaround for detector hardware changes
 - Provide quick feedback on data quality
- D0 MC generation uses remote resources
 - Non-Grid, Open Science Grid, Native SamGrid
 - MC generation prioritized by physics analysis needs